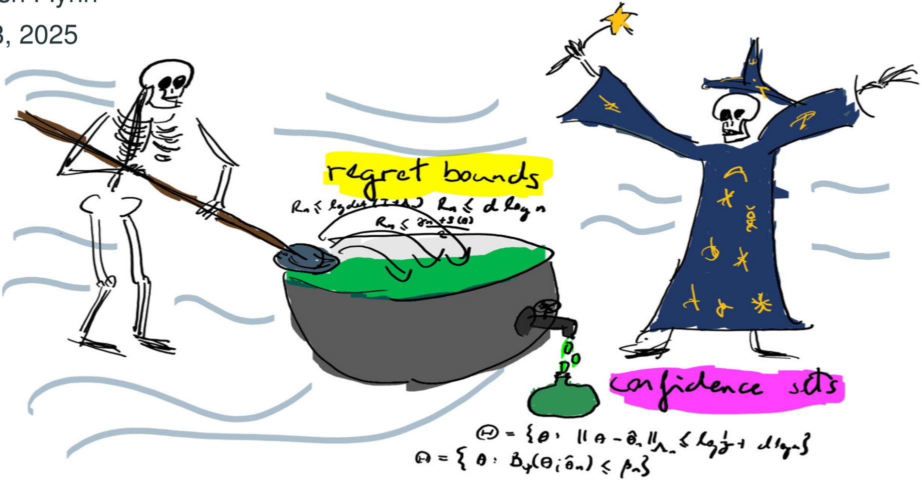
Confidence Sequences for Generalised Linear Models via Regret Analysis

Hamish Flynn

July 3, 2025



Warmest Thanks

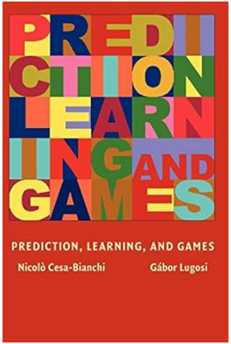
Eugenio Clerico

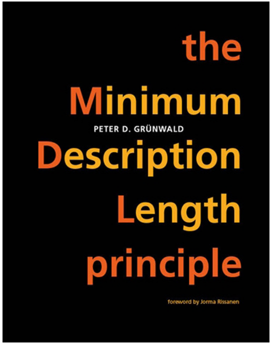
Wojciech Kotłowski

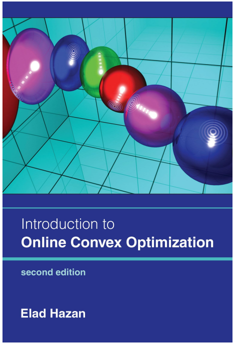
Gergely Neu

This Work

Assume we know a bit about online learning (or universal coding).

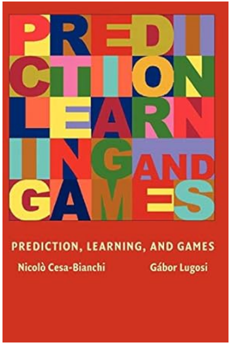


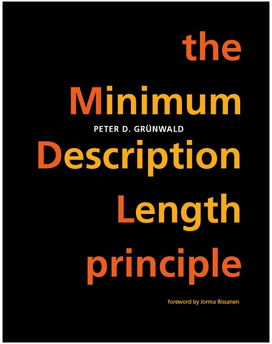


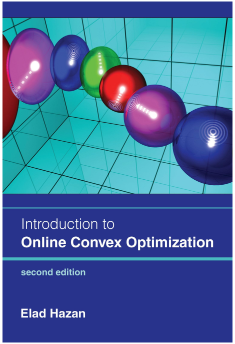


This Work

Assume we know a bit about online learning (or universal coding).







We want to construct confidence sequences for GLMs without doing any actual work.

Generalised Linear Models

Generalised Linear Model:

- Covariates $X_1, \ldots, X_n \in \mathbb{R}^d$
- Responses $Y_1, \ldots, Y_n \in \mathbb{R}$
- Likelihood $p(Y_t | X_t, \theta^\star) = \exp\big(\langle \theta^\star, x \rangle y - \psi(\langle \theta^\star, x \rangle)\big) h(y)$

The log-partition function $\psi : \mathbb{R} \to \mathbb{R}$ is convex.

Generalised Linear Model:

- Covariates $X_1, \dots, X_n \in \mathbb{R}^d$
- Responses $Y_1, \dots, Y_n \in \mathbb{R}$
- Likelihood $p(Y_t | X_t, \theta^\star) = \exp\big(\langle \theta^\star, x\rangle y - \psi(\langle \theta^\star, x\rangle)\big) h(y)$

The log-partition function $\psi : \mathbb{R} \to \mathbb{R}$ is convex.

Log-Likelihood Loss: Define $\ell_t(\theta) = -\log(p(Y_t | X_t, \theta))$.

Generalised Linear Models

Generalised Linear Model:

- Covariates $X_1, \dots, X_n \in \mathbb{R}^d$
- Responses $Y_1, \dots, Y_n \in \mathbb{R}$
- Likelihood $p(Y_t|X_t, \theta^\star) = \exp\big(\langle \theta^\star, x\rangle y - \psi(\langle \theta^\star, x\rangle)\big) h(y)$

The log-partition function $\psi : \mathbb{R} \to \mathbb{R}$ is convex.

Log-Likelihood Loss: Define $\ell_t(\theta) = -\log(p(Y_t|X_t, \theta))$.

Adaptive Design: X_t depends on $X_1, Y_1, \dots, X_{t-1}, Y_{t-1}$.

Generalised Linear Models

Generalised Linear Model:

- Covariates $X_1, \ldots, X_n \in \mathbb{R}^d$
- Responses $Y_1, \ldots, Y_n \in \mathbb{R}$
- Likelihood $p(Y_t | X_t, \theta^\star) = \exp\big(\langle \theta^\star, x\rangle y - \psi(\langle \theta^\star, x\rangle)\big) h(y)$

The log-partition function $\psi : \mathbb{R} \to \mathbb{R}$ is convex.

Log-Likelihood Loss: Define $\ell_t(\theta) = -\log(p(Y_t | X_t, \theta))$.

Adaptive Design: X_t depends on $X_1, Y_1, \ldots, X_{t-1}, Y_{t-1}$.

Oblivious Design: X_t does not depend on $Y_1, \ldots, Y_{t-1}$.

Objective and Claim

Adaptive Design: For $\delta \in (0, 1]$, a δ-*confidence sequence* for $\theta^\star$ is a sequence of sets $\Theta_1, \Theta_2, \ldots$, such that

$$\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta \,.$$

Objective and Claim

Adaptive Design: For $\delta \in (0, 1]$, a δ-*confidence sequence* for $\theta^\star$ is a sequence of sets $\Theta_1, \Theta_2, \dots$, such that

$$\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta\,.$$

Oblivious Design: A δ-*confidence set* for $\theta^\star$ is a set Θ_n, such that

$$\mathbb{P}(\theta^\star \notin \Theta_n) \leq \delta\,.$$

Objective and Claim

Adaptive Design: For $\delta \in (0, 1]$, a δ-*confidence sequence* for $\theta^\star$ is a sequence of sets $\Theta_1, \Theta_2, \ldots$, such that

$$\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta\,.$$

Oblivious Design: A δ-*confidence set* for $\theta^\star$ is a set Θ_n, such that

$$\mathbb{P}(\theta^\star \notin \Theta_n) \leq \delta\,.$$

In this talk, we will (mostly) focus on sets of the form

$$\Theta_n = \left\{ \theta : \sum_{t=1}^{n} \ell_t(\theta) - \inf_{\theta' \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(\theta') \leq \beta_n \right\}$$

Objective and Claim

Adaptive Design: For $\delta \in (0, 1]$, a δ-*confidence sequence* for $\theta^\star$ is a sequence of sets $\Theta_1, \Theta_2, \ldots$, such that

$$\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta\,.$$

Oblivious Design: A δ-*confidence set* for $\theta^\star$ is a set Θ_n, such that

$$\mathbb{P}(\theta^\star \notin \Theta_n) \leq \delta\,.$$

In this talk, we will (mostly) focus on sets of the form

$$\Theta_n = \left\{ \theta : \sum_{t=1}^{n} \ell_t(\theta) - \inf_{\theta' \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(\theta') \leq \beta_n \right\}$$

Online-to-confidence-set conversion: Use the output and/or regret bound of an online learning algorithm to determine β_n.

Objective and Claim

Adaptive Design: For $\delta \in (0,1]$, a δ-*confidence sequence* for $\theta^\star$ is a sequence of sets $\Theta_1, \Theta_2, \ldots$, such that

$$\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta \,.$$

Oblivious Design: A δ-*confidence set* for $\theta^\star$ is a set Θ_n, such that

$$\mathbb{P}(\theta^\star \notin \Theta_n) \leq \delta \,.$$

In this talk, we will (mostly) focus on sets of the form

$$\Theta_n = \left\{ \theta : \sum_{t=1}^{n} \ell_t(\theta) - \inf_{\theta' \in \mathbb{R}^d} \sum_{t=1}^{n} \ell_t(\theta') \leq \beta_n \right\}$$

Online-to-confidence-set conversion: Use the output and/or regret bound of an online learning algorithm to determine β_n.

Claim: We can recover all confidence sequences for GLMs via OTCS (at least all confidence sequences with non-asymptotic coverage guarantees).

Online-To-Confidence-Set Conversion

(for adaptive design)

Sequential Probability Assignment

Protocol: For $t = 1, 2, \ldots, n$:

1. Environment reveals X_t to the learner
2. Learner picks $Q_t \in \Delta_\Theta$ with density q_t
3. Environment reveals Y_t to the learner,
4. Learner incurs the log loss $\mathcal{L}_t(q_t) = -\log \int \exp(-\ell_t(\theta)) q_t(\theta) \, \mathrm{d}\theta$

Sequential Probability Assignment

Protocol: For $t = 1, 2, \dots, n$:

1. Environment reveals X_t to the learner
2. Learner picks $Q_t \in \Delta_\Theta$ with density q_t
3. Environment reveals Y_t to the learner,
4. Learner incurs the log loss $\mathcal{L}_t(q_t) = -\log \int \exp(-\ell_t(\theta)) q_t(\theta) \, \mathrm{d}\theta$

$q^n = (q_1, \dots, q_n)$ must be predictable w.r.t. $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$, where $\mathcal{F}_t = \sigma(X_1, Y_1, \dots, X_t, Y_t, X_{t+1})$.

Sequential Probability Assignment

Protocol: For $t = 1, 2, \ldots, n$:

1. Environment reveals X_t to the learner
2. Learner picks $Q_t \in \Delta_\Theta$ with density q_t
3. Environment reveals Y_t to the learner,
4. Learner incurs the log loss $\mathcal{L}_t(q_t) = -\log \int \exp(-\ell_t(\theta)) q_t(\theta) \, \mathrm{d}\theta$

$q^n = (q_1, \ldots, q_n)$ must be predictable w.r.t. $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$, where $\mathcal{F}_t = \sigma(X_1, Y_1, \ldots, X_t, Y_t, X_{t+1})$.

Regret: The regret of q^n w.r.t. a comparator $\bar{\theta} \in \Theta$ is

$$\mathrm{regret}_{q^n}(\bar{\theta}) = \sum_{t=1}^{n} \left(\mathcal{L}_t(q_t) - \ell_t(\bar{\theta}) \right) .$$

Sequential Probability Assignment

Protocol: For $t = 1, 2, \ldots, n$:

1. Environment reveals X_t to the learner
2. Learner picks $Q_t \in \Delta_\Theta$ with density q_t
3. Environment reveals Y_t to the learner,
4. Learner incurs the log loss $\mathcal{L}_t(q_t) = -\log \int \exp(-\ell_t(\theta)) q_t(\theta) \,\mathrm{d}\theta$

$q^n = (q_1, \ldots, q_n)$ must be predictable w.r.t. $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$, where $\mathcal{F}_t = \sigma(X_1, Y_1, \ldots, X_t, Y_t, X_{t+1})$.

Regret: The regret of q^n w.r.t. a comparator $\bar{\theta} \in \Theta$ is

$$\mathrm{regret}_{q^n}(\bar{\theta}) = \sum_{t=1}^{n} \left(\mathcal{L}_t(q_t) - \ell_t(\bar{\theta}) \right) .$$

The minimax regret $\inf_{q^n} \sup_{\bar{\theta}} \mathrm{regret}_{q^n}(\bar{\theta})$ for linear models (and some GLMs) is of the order $d \log(n)$.

Sequential Probability Assignment

Protocol: For $t = 1, 2, \ldots, n$:

1. Environment reveals X_t to the learner
2. Learner picks $Q_t \in \Delta_\Theta$ with density q_t
3. Environment reveals Y_t to the learner,
4. Learner incurs the log loss $\mathcal{L}_t(q_t) = -\log \int \exp(-\ell_t(\theta)) q_t(\theta) \, \mathrm{d}\theta$

$q^n = (q_1, \ldots, q_n)$ must be predictable w.r.t. $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$, where $\mathcal{F}_t = \sigma(X_1, Y_1, \ldots, X_t, Y_t, X_{t+1})$.

Regret: The regret of q^n w.r.t. a comparator $\bar{\theta} \in \Theta$ is

$$\mathrm{regret}_{q^n}(\bar{\theta}) = \sum_{t=1}^{n} \left(\mathcal{L}_t(q_t) - \ell_t(\bar{\theta}) \right) .$$

The minimax regret $\inf_{q^n} \sup_{\bar{\theta}} \mathrm{regret}_{q^n}(\bar{\theta})$ for linear models (and some GLMs) is of the order $d \log(n)$.

Note: This can be made more general by playing distributions on $\mathcal{Y}$ (see our paper).

Online-To-Confidence Set Conversion (Adaptive Design)

For any sequence of comparators $\bar{\theta}_1, \bar{\theta}_2, \ldots$ and any $\mathbb{F}$-predictable q^n, the sets $\Theta_1, \Theta_2, \ldots$ form a δ-CS, where

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \left(\ell_t(\theta) - \ell_t(\bar{\theta}_n) \right) \leq \text{regret}_{q^n}(\bar{\theta}_n) + \log \frac{1}{\delta} \right\} .$$

Online-To-Confidence Set Conversion (Adaptive Design)

For any sequence of comparators $\bar{\theta}_1, \bar{\theta}_2, \ldots$ and any $\mathbb{F}$-predictable q^n, the sets $\Theta_1, \Theta_2, \ldots$ form a δ-CS, where

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \left(\ell_t(\theta) - \ell_t(\bar{\theta}_n) \right) \leq \mathrm{regret}_{q^n}(\bar{\theta}_n) + \log \tfrac{1}{\delta} \right\} .$$

Proof. First,

$$\sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \ell_t(\bar{\theta}_n) \right) = \sum_{t=1}^{n} \left(\mathcal{L}_t(q_t) - \ell_t(\bar{\theta}_n) \right) + \sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \mathcal{L}_t(q_t) \right) .$$

Online-To-Confidence Set Conversion (Adaptive Design)

For any sequence of comparators $\bar{\theta}_1, \bar{\theta}_2, \ldots$ and any $\mathbb{F}$-predictable q^n, the sets $\Theta_1, \Theta_2, \ldots$ form a δ-CS, where

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \left(\ell_t(\theta) - \ell_t(\bar{\theta}_n) \right) \leq \mathrm{regret}_{q^n}(\bar{\theta}_n) + \log \tfrac{1}{\delta} \right\} .$$

Proof. First,

$$\sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \ell_t(\bar{\theta}_n) \right) = \underbrace{\sum_{t=1}^{n} \left(\mathcal{L}_t(q_t) - \ell_t(\bar{\theta}_n) \right)}_{\mathrm{regret}_{q^n}(\bar{\theta}_n)} + \sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \mathcal{L}_t(q_t) \right) .$$

Online-To-Confidence Set Conversion (Adaptive Design)

For any sequence of comparators $\bar{\theta}_1, \bar{\theta}_2, \ldots$ and any $\mathbb{F}$-predictable q^n, the sets $\Theta_1, \Theta_2, \ldots$ form a δ-CS, where

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \left(\ell_t(\theta) - \ell_t(\bar{\theta}_n) \right) \leq \operatorname{regret}_{q^n}(\bar{\theta}_n) + \log \tfrac{1}{\delta} \right\} .$$

Proof. First,

$$\sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \ell_t(\bar{\theta}_n) \right) = \underbrace{\sum_{t=1}^{n} \left(\mathcal{L}_t(q_t) - \ell_t(\bar{\theta}_n) \right)}_{\operatorname{regret}_{q^n}(\bar{\theta}_n)} + \sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \mathcal{L}_t(q_t) \right) .$$

Next, we notice that the second term is the logarithm of a non-negative $\mathbb{F}$-martingale.

$$\exp\left(\sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \mathcal{L}_t(q_t) \right) \right) = \prod_{t=1}^{n} \int \frac{p(Y_t | X_t, \theta)}{p(Y_t | X_t, \theta^\star)} q_t(\theta) \, \mathrm{d}\theta \, .$$

Online-To-Confidence Set Conversion (Adaptive Design)

For any sequence of comparators $\bar{\theta}_1, \bar{\theta}_2, \ldots$ and any $\mathbb{F}$-predictable q^n, the sets $\Theta_1, \Theta_2, \ldots$ form a δ-CS, where

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \left(\ell_t(\theta) - \ell_t(\bar{\theta}_n)\right) \leq \mathrm{regret}_{q^n}(\bar{\theta}_n) + \log \tfrac{1}{\delta} \right\} .$$

Proof. First,

$$\sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \ell_t(\bar{\theta}_n)\right) = \underbrace{\sum_{t=1}^{n} \left(\mathcal{L}_t(q_t) - \ell_t(\bar{\theta}_n)\right)}_{\mathrm{regret}_{q^n}(\bar{\theta}_n)} + \sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \mathcal{L}_t(q_t)\right) .$$

Next, we notice that the second term is the logarithm of a non-negative $\mathbb{F}$-martingale.

$$\exp\left(\sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \mathcal{L}_t(q_t)\right) \right) = \prod_{t=1}^{n} \int \frac{p(Y_t | X_t, \theta)}{p(Y_t | X_t, \theta^\star)} q_t(\theta) \, \mathrm{d}\theta .$$

Therefore,

$$\mathbb{P}\left(\exists n \geq 1, \sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \ell_t(\bar{\theta}_n)\right) \geq \mathrm{regret}_{q^n}(\bar{\theta}_n) + \log \tfrac{1}{\delta} \right) \leq \delta .$$

Online-To-Confidence-Set Conversion

(for oblivious design)

Transductive Sequential Probability Assignment

Protocol: The environment reveals $X_1, \ldots, X_n \in \mathbb{R}^d$. For $t = 1, 2, \ldots, n$:

1. Learner picks $Q_t \in \Delta_\Theta$ with density q_t
2. Environment reveals $Y_t \in \mathbb{R}$ to the learner,
3. Learner incurs the log loss $\mathcal{L}_t(q_t) = -\log \int \exp(-\ell_t(\theta)) q_t(\theta) \, \mathrm{d}\theta$

Transductive Sequential Probability Assignment

Protocol: The environment reveals $X_1, \ldots, X_n \in \mathbb{R}^d$. For $t = 1, 2, \ldots, n$:

1. Learner picks $Q_t \in \Delta_\Theta$ with density q_t
2. Environment reveals $Y_t \in \mathbb{R}$ to the learner,
3. Learner incurs the log loss $\mathcal{L}_t(q_t) = -\log \int \exp(-\ell_t(\theta)) q_t(\theta) \, \mathrm{d}\theta$

q^n must be predictable w.r.t. $\widetilde{\mathbb{F}} = (\widetilde{\mathcal{F}}_t)_{t=0}^n$, where $\widetilde{\mathcal{F}}_t = \sigma(X_1, \ldots, X_n, Y_1, \ldots, Y_t)$.

Transductive Sequential Probability Assignment

Protocol: The environment reveals $X_1, \ldots, X_n \in \mathbb{R}^d$. For $t = 1, 2, \ldots, n$:

1. Learner picks $Q_t \in \Delta_\Theta$ with density q_t
2. Environment reveals $Y_t \in \mathbb{R}$ to the learner,
3. Learner incurs the log loss $\mathcal{L}_t(q_t) = -\log \int \exp(-\ell_t(\theta))q_t(\theta)\, \mathrm{d}\theta$

q^n must be predictable w.r.t. $\widetilde{\mathbb{F}} = (\widetilde{\mathcal{F}}_t)_{t=0}^n$, where $\widetilde{\mathcal{F}}_t = \sigma(X_1, \ldots, X_n, Y_1, \ldots, Y_t)$.

Regret: The regret of $q^n = (q_1, \ldots, q_n)$ w.r.t. a comparator $\bar{\theta} \in \Theta$ is

$$\mathrm{regret}_{q^n}(\bar{\theta}) = \sum_{t=1}^{n} \left(\mathcal{L}_t(q_t) - \ell_t(\bar{\theta})\right).$$

Online-To-Confidence Set Conversion (Oblivious Design)

For any comparator $\bar{\theta}_n$ and any $\widetilde{\mathbb{F}}$-predictable q^n, the set Θ_n is a δ-CS, where

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \big(\ell_t(\theta) - \ell_t(\bar{\theta}_n) \big) \leq \mathrm{regret}_{q^n}(\bar{\theta}_n) + \log \tfrac{1}{\delta} \right\} .$$

Online-To-Confidence Set Conversion (Oblivious Design)

For any comparator $\bar{\theta}_n$ and any $\widetilde{\mathbb{F}}$-predictable q^n, the set Θ_n is a δ-CS, where

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \big(\ell_t(\theta) - \ell_t(\bar{\theta}_n)\big) \leq \mathrm{regret}_{q^n}(\bar{\theta}_n) + \log \tfrac{1}{\delta} \right\} .$$

Proof. Basically the same as last time. First,

$$\sum_{t=1}^{n} \big(\ell_t(\theta^\star) - \ell_t(\bar{\theta}_n)\big) = \mathrm{regret}_{q^n}(\bar{\theta}_n) + \sum_{t=1}^{n} \big(\ell_t(\theta^\star) - \mathcal{L}_t(q_t)\big) .$$

Online-To-Confidence Set Conversion (Oblivious Design)

For any comparator $\bar{\theta}_n$ and any $\widetilde{\mathbb{F}}$-predictable q^n, the set Θ_n is a δ-CS, where

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \big(\ell_t(\theta) - \ell_t(\bar{\theta}_n)\big) \leq \mathrm{regret}_{q^n}(\bar{\theta}_n) + \log \tfrac{1}{\delta} \right\} .$$

Proof. Basically the same as last time. First,

$$\sum_{t=1}^{n} \big(\ell_t(\theta^\star) - \ell_t(\bar{\theta}_n)\big) = \mathrm{regret}_{q^n}(\bar{\theta}_n) + \sum_{t=1}^{n} \big(\ell_t(\theta^\star) - \mathcal{L}_t(q_t)\big) .$$

As long as the design is oblivious, the second term is the logarithm of a non-negative $\widetilde{\mathbb{F}}$-martingale. Therefore,

$$\mathbb{P}\left(\sum_{t=1}^{n} \big(\ell_t(\theta^\star) - \ell_t(\bar{\theta}_n)\big) \geq \mathrm{regret}_{q^n}(\bar{\theta}_n) + \log \tfrac{1}{\delta} \right) \leq \delta .$$

Online-To-Confidence-Set Conversions for Smooth GLMs

Bregman Divergence and Bregman Information Gain

For a convex differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, the *Bregman divergence* is

$$\mathcal{B}_f(\theta, \theta') = f(\theta) - f(\theta') - \langle \theta - \theta', \nabla f(\theta') \rangle \,.$$

Bregman Divergence and Bregman Information Gain

For a convex differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, the *Bregman divergence* is

$$\mathcal{B}_f(\theta, \theta') = f(\theta) - f(\theta') - \langle \theta - \theta', \nabla f(\theta') \rangle \,.$$

For a convex differentiable function $\rho : \mathbb{R}^d \to \mathbb{R}$, let $Z_n^\rho(\theta) = \sum_{t=1}^n \ell_t(\theta) + \rho(\theta)$ and $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \mathbb{R}^d} Z_n^\rho(\theta)$.

Bregman Divergence and Bregman Information Gain

For a convex differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, the *Bregman divergence* is

$$\mathcal{B}_f(\theta, \theta') = f(\theta) - f(\theta') - \langle \theta - \theta', \nabla f(\theta') \rangle \, .$$

For a convex differentiable function $\rho : \mathbb{R}^d \to \mathbb{R}$, let $Z_n^\rho(\theta) = \sum_{t=1}^n \ell_t(\theta) + \rho(\theta)$ and $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \mathbb{R}^d} Z_n^\rho(\theta)$.

The *Bregman information gain* is

$$\gamma_n^\rho = -\log \left(\frac{\int \exp(-\mathcal{B}_{Z_n^\rho}(\theta, \widehat{\theta}_n)) \mathrm{d}\theta}{\int \exp(-\rho(\theta)) \mathrm{d}\theta} \right) \, .$$

Bregman Divergence and Bregman Information Gain

For a convex differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, the *Bregman divergence* is

$$\mathcal{B}_f(\theta, \theta') = f(\theta) - f(\theta') - \langle \theta - \theta', \nabla f(\theta') \rangle \,.$$

For a convex differentiable function $\rho : \mathbb{R}^d \to \mathbb{R}$, let $Z_n^\rho(\theta) = \sum_{t=1}^n \ell_t(\theta) + \rho(\theta)$ and $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \mathbb{R}^d} Z_n^\rho(\theta)$.

The *Bregman information gain* is

$$\gamma_n^\rho = -\log\left(\frac{\int \exp(-\mathcal{B}_{Z_n^\rho}(\theta, \widehat{\theta}_n))\mathrm{d}\theta}{\int \exp(-\rho(\theta))\mathrm{d}\theta}\right) \,.$$

If ψ is M-smooth ($|\psi''(z)| \leq M$) and $\rho = \frac{1}{2\gamma^2}\|\theta\|_2^2$, then

$$\gamma_n^\rho \leq \frac{1}{2}\log\det(M\gamma^2\Lambda_n + \mathrm{Id}) \leq \frac{d}{2}\log(1 + \tfrac{\gamma^2 M L^2 n}{d}) \,,$$

where $\Lambda_n = \sum_{t=1}^n X_t X_t^\top$ and $L = \max_{t\in[n]} \|X_t\|_2$.

Exponentially Weighted Average Forecaster

The Exponentially Weighted Average (EWA) forecaster takes as input a prior

$$q_1(\theta) \propto \exp(-\rho(\theta)) .$$

Exponentially Weighted Average Forecaster

The Exponentially Weighted Average (EWA) forecaster takes as input a prior

$$q_1(\theta) \propto \exp(-\rho(\theta)) \,.$$

For subsequent rounds $t = 2, 3, \ldots$, the EWA forecaster plays

$$q_t(\theta) \propto q_1(\theta) \exp\left(\sum_{k=1}^{t-1} \ell_k(\theta)\right) \,.$$

Exponentially Weighted Average Forecaster

The Exponentially Weighted Average (EWA) forecaster takes as input a prior

$$q_1(\theta) \propto \exp(-\rho(\theta)) .$$

For subsequent rounds $t = 2, 3, \ldots$, the EWA forecaster plays

$$q_t(\theta) \propto q_1(\theta) \exp\left(\sum_{k=1}^{t-1} \ell_k(\theta)\right) .$$

Claim: For any choice of ρ,

$$\text{regret}_{q^n}(\bar{\theta}_n) \leq \gamma_n^\rho + \rho(\bar{\theta}_n) .$$

OTCS for Smooth GLMs (Adaptive Design)

Suppose that ψ is M-smooth, and fix $\gamma > 0$. Set $\rho = \frac{1}{2\gamma^2}\|\theta\|_2^2$ and let $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \mathbb{R}^d}\{\sum_{t=1}^n \ell_t(\theta) + \rho(\theta)\}$.

OTCS for Smooth GLMs (Adaptive Design)

Suppose that ψ is M-smooth, and fix $\gamma > 0$. Set $\rho = \frac{1}{2\gamma^2}\|\theta\|_2^2$ and let $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \mathbb{R}^d}\{\sum_{t=1}^n \ell_t(\theta) + \rho(\theta)\}$.

Then, for any $\delta \in (0, 1]$, the sets $\Theta_1, \Theta_2, \ldots$ satisfy $\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta$, where

$$\Theta_n = \left\{\theta : \sum_{t=1}^n \ell_t(\theta) - \sum_{t=1}^n \ell_t(\widehat{\theta}_n) \leq \frac{1}{2}\log\det(\gamma^2 M \Lambda_n + \mathrm{Id}) + \frac{\|\widehat{\theta}_n\|_2^2}{2\gamma^2} + \log\frac{1}{\delta}\right\}$$

OTCS for Smooth GLMs (Adaptive Design)

Suppose that ψ is M-smooth, and fix $\gamma > 0$. Set $\rho = \frac{1}{2\gamma^2}\|\theta\|_2^2$ and let $\widehat{\theta}_n = \mathrm{argmin}_{\theta \in \mathbb{R}^d}\{\sum_{t=1}^n \ell_t(\theta) + \rho(\theta)\}$.

Then, for any $\delta \in (0, 1]$, the sets $\Theta_1, \Theta_2, \ldots$ satisfy $\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta$, where

$$\Theta_n = \left\{\theta : \sum_{t=1}^n \ell_t(\theta) - \sum_{t=1}^n \ell_t(\widehat{\theta}_n) \leq \frac{1}{2}\log\det(\gamma^2 M \Lambda_n + \mathrm{Id}) + \frac{\|\widehat{\theta}_n\|_2^2}{2\gamma^2} + \log\frac{1}{\delta}\right\}$$

If it is known that $\|\theta^\star\|_2 \leq B$, then we can also use

$$\Theta_n = \left\{\theta : \mathcal{B}_{Z_n^\rho}(\theta, \widehat{\theta}_n) \leq \frac{1}{2}\log\det(\gamma^2 M \Lambda_n + \mathrm{Id}) + \frac{B^2}{2\gamma^2} + \log\frac{1}{\delta}\right\}$$

OTCS for Smooth GLMs (Adaptive Design)

Suppose that ψ is M-smooth, and fix $\gamma > 0$. Set $\rho = \frac{1}{2\gamma^2}\|\theta\|_2^2$ and let $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \mathbb{R}^d}\{\sum_{t=1}^n \ell_t(\theta) + \rho(\theta)\}$.

Then, for any $\delta \in (0, 1]$, the sets $\Theta_1, \Theta_2, \ldots$ satisfy $\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta$, where

$$\Theta_n = \left\{\theta : \sum_{t=1}^n \ell_t(\theta) - \sum_{t=1}^n \ell_t(\widehat{\theta}_n) \leq \tfrac{1}{2}\log\det(\gamma^2 M \Lambda_n + \mathrm{Id}) + \frac{\|\widehat{\theta}_n\|_2^2}{2\gamma^2} + \log\tfrac{1}{\delta}\right\}$$

If it is known that $\|\theta^\star\|_2 \leq B$, then we can also use

$$\Theta_n = \left\{\theta : \mathcal{B}_{Z_n^\rho}(\theta, \widehat{\theta}_n) \leq \tfrac{1}{2}\log\det(\gamma^2 M \Lambda_n + \mathrm{Id}) + \frac{B^2}{2\gamma^2} + \log\tfrac{1}{\delta}\right\}$$

For general (smooth) GLMs, we match the best confidence sequence that we're aware of (except the radius of ours is dimension-free).

OTCS for Smooth GLMs (Adaptive Design)

Suppose that ψ is M-smooth, and fix $\gamma > 0$. Set $\rho = \frac{1}{2\gamma^2}\|\theta\|_2^2$ and let $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \mathbb{R}^d}\{\sum_{t=1}^n \ell_t(\theta) + \rho(\theta)\}$.

Then, for any $\delta \in (0,1]$, the sets $\Theta_1, \Theta_2, \ldots$ satisfy $\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta$, where

$$\Theta_n = \left\{\theta : \sum_{t=1}^n \ell_t(\theta) - \sum_{t=1}^n \ell_t(\widehat{\theta}_n) \leq \frac{1}{2}\log\det(\gamma^2 M \Lambda_n + \mathrm{Id}) + \frac{\|\widehat{\theta}_n\|_2^2}{2\gamma^2} + \log\frac{1}{\delta}\right\}$$

If it is known that $\|\theta^\star\|_2 \leq B$, then we can also use

$$\Theta_n = \left\{\theta : \mathcal{B}_{Z_n^\rho}(\theta, \widehat{\theta}_n) \leq \frac{1}{2}\log\det(\gamma^2 M \Lambda_n + \mathrm{Id}) + \frac{B^2}{2\gamma^2} + \log\frac{1}{\delta}\right\}$$

For general (smooth) GLMs, we match the best confidence sequence that we're aware of (except the radius of ours is dimension-free).

For linear models, the Bregman ball becomes the ellipsoid,

$$\Theta_n = \left\{\theta : \|\theta - \widehat{\theta}_n\|^2_{\Lambda_n + \frac{1}{\gamma^2}\mathrm{Id}} \leq \log\det(\gamma^2 M \Lambda_n + \mathrm{Id}) + \frac{B^2}{\gamma^2} + 2\log\frac{1}{\delta}\right\}$$

OTCS for Smooth GLMs (Adaptive Design)

Suppose that ψ is M-smooth, and fix $\gamma > 0$. Set $\rho = \frac{1}{2\gamma^2}\|\theta\|_2^2$ and let $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \mathbb{R}^d}\{\sum_{t=1}^n \ell_t(\theta) + \rho(\theta)\}$.

Then, for any $\delta \in (0, 1]$, the sets $\Theta_1, \Theta_2, \ldots$ satisfy $\mathbb{P}(\exists n \geq 1 : \theta^\star \notin \Theta_n) \leq \delta$, where

$$\Theta_n = \left\{\theta : \sum_{t=1}^n \ell_t(\theta) - \sum_{t=1}^n \ell_t(\widehat{\theta}_n) \leq \tfrac{1}{2}\log\det(\gamma^2 M\Lambda_n + \mathrm{Id}) + \frac{\|\widehat{\theta}_n\|_2^2}{2\gamma^2} + \log\tfrac{1}{\delta}\right\}$$

If it is known that $\|\theta^\star\|_2 \leq B$, then we can also use

$$\Theta_n = \left\{\theta : \mathcal{B}_{Z_n^\rho}(\theta, \widehat{\theta}_n) \leq \tfrac{1}{2}\log\det(\gamma^2 M\Lambda_n + \mathrm{Id}) + \frac{B^2}{2\gamma^2} + \log\tfrac{1}{\delta}\right\}$$

For general (smooth) GLMs, we match the best confidence sequence that we're aware of (except the radius of ours is dimension-free).

For linear models, the Bregman ball becomes the ellipsoid,

$$\Theta_n = \left\{\theta : \|\theta - \widehat{\theta}_n\|^2_{\Lambda_n + \frac{1}{\gamma^2}\mathrm{Id}} \leq \log\det(\gamma^2 M\Lambda_n + \mathrm{Id}) + \frac{B^2}{\gamma^2} + 2\log\tfrac{1}{\delta}\right\}$$

This matches the one you would get from self-normalised concentration (with slightly better constants).

OTCS for Oblivious Design

Let $\mathcal{S}_{n,b} = \{\theta : \max_{t\in[n]} |\langle\theta, X_t\rangle| \leq b\}$, let $\widehat{\theta}_{n,b} = \operatorname{argmin}_{\theta\in\mathcal{S}_{n,b}} \sum_{t=1}^n \ell_t(\theta)$ and let $\Psi(\theta) = \sum_{t=1}^n \psi(\langle\theta, X_t\rangle)$. Suppose that $\theta^\star$ satisfies $\sup_{t\in[n]} |\langle\theta^\star, X_t\rangle| \leq b$ (w.p. 1) that ψ is M-smooth on $\mathbb{R}$ and m-strongly-convex on $[-b, b]$. Set $\rho(\theta) = \frac{1}{2\gamma^2}\|\theta - \theta^\star\|_{\Lambda_n}^2$.

OTCS for Oblivious Design

Let $\mathcal{S}_{n,b} = \{\theta : \max_{t\in[n]} |\langle \theta, X_t\rangle| \le b\}$, let $\widehat{\theta}_{n,b} = \operatorname{argmin}_{\theta\in\mathcal{S}_{n,b}} \sum_{t=1}^n \ell_t(\theta)$ and let $\Psi(\theta) = \sum_{t=1}^n \psi(\langle\theta, X_t\rangle)$. Suppose that $\theta^\star$ satisfies $\sup_{t\in[n]} |\langle\theta^\star, X_t\rangle| \le b$ (w.p. 1) that ψ is M-smooth on $\mathbb{R}$ and m-strongly-convex on $[-b, b]$. Set $\rho(\theta) = \frac{1}{2\gamma^2}\|\theta - \theta^\star\|_{\Lambda_n}^2$.

For any $\delta \in (0, 1]$, the OTCS with oblivious design tells us that

$$\mathbb{P}\left(\sum_{t=1}^n \ell_t(\theta^\star) - \sum_{t=1}^n \ell_t(\widehat{\theta}_{n,b}) \le \frac{d}{2}\log(1 + M\gamma^2) + \frac{1}{2\gamma^2}\|\widehat{\theta}_{n,b} - \theta^\star\|_{\Lambda_n}^2 + \log\frac{1}{\delta}\right).$$

OTCS for Oblivious Design

Let $\mathcal{S}_{n,b} = \{\theta : \max_{t\in[n]} |\langle \theta, X_t \rangle| \le b\}$, let $\widehat{\theta}_{n,b} = \operatorname{argmin}_{\theta \in \mathcal{S}_{n,b}} \sum_{t=1}^n \ell_t(\theta)$ and let $\Psi(\theta) = \sum_{t=1}^n \psi(\langle \theta, X_t \rangle)$. Suppose that $\theta^\star$ satisfies $\sup_{t\in[n]} |\langle \theta^\star, X_t \rangle| \le b$ (w.p. 1) that ψ is M-smooth on $\mathbb{R}$ and m-strongly-convex on $[-b, b]$. Set $\rho(\theta) = \frac{1}{2\gamma^2} \|\theta - \theta^\star\|_{\Lambda_n}^2$.

For any $\delta \in (0, 1]$, the OTCS with oblivious design tells us that

$$\mathbb{P}\left(\sum_{t=1}^n \ell_t(\theta^\star) - \sum_{t=1}^n \ell_t(\widehat{\theta}_{n,b}) \le \frac{d}{2} \log(1 + M\gamma^2) + \frac{1}{2\gamma^2} \|\widehat{\theta}_{n,b} - \theta^\star\|_{\Lambda_n}^2 + \log \frac{1}{\delta} \right) .$$

Using the first-order optimality condition satisfied by $\widehat{\theta}_{n,b}$ and strong convexity of ψ on $[-b, b]$,

$$\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) \le \sum_{t=1}^n \ell_t(\theta) - \sum_{t=1}^n \ell_t(\widehat{\theta}_{n,b}) , \quad \text{and} \quad \frac{1}{2\gamma^2} \|\widehat{\theta}_{n,b} - \theta^\star\|_{\Lambda_n}^2 \le \frac{1}{m\gamma^2} \mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) .$$

OTCS for Oblivious Design

Let $\mathcal{S}_{n,b} = \{\theta : \max_{t\in[n]} |\langle\theta, X_t\rangle| \le b\}$, let $\widehat{\theta}_{n,b} = \operatorname{argmin}_{\theta\in\mathcal{S}_{n,b}} \sum_{t=1}^{n} \ell_t(\theta)$ and let $\Psi(\theta) = \sum_{t=1}^{n} \psi(\langle\theta, X_t\rangle)$. Suppose that $\theta^\star$ satisfies $\sup_{t\in[n]} |\langle\theta^\star, X_t\rangle| \le b$ (w.p. 1) that ψ is M-smooth on $\mathbb{R}$ and m-strongly-convex on $[-b, b]$. Set $\rho(\theta) = \frac{1}{2\gamma^2}\|\theta - \theta^\star\|^2_{\Lambda_n}$.

For any $\delta \in (0, 1]$, the OTCS with oblivious design tells us that

$$\mathbb{P}\left(\sum_{t=1}^{n}\ell_t(\theta^\star) - \sum_{t=1}^{n}\ell_t(\widehat{\theta}_{n,b}) \le \frac{d}{2}\log(1 + M\gamma^2) + \frac{1}{2\gamma^2}\|\widehat{\theta}_{n,b} - \theta^\star\|^2_{\Lambda_n} + \log\frac{1}{\delta}\right) .$$

Using the first-order optimality condition satisfied by $\widehat{\theta}_{n,b}$ and strong convexity of ψ on $[-b, b]$,

$$\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) \le \sum_{t=1}^{n}\ell_t(\theta) - \sum_{t=1}^{n}\ell_t(\widehat{\theta}_{n,b}) \,, \quad \text{and} \quad \frac{1}{2\gamma^2}\|\widehat{\theta}_{n,b} - \theta^\star\|^2_{\Lambda_n} \le \frac{1}{m\gamma^2}\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) .$$

Therefore (with $\gamma^2 = 2/m$), the set Θ_n satisfies $\mathbb{P}(\theta^\star \notin \Theta_n) \le \delta$, where

$$\Theta_n = \left\{\theta : \mathcal{B}_\Psi(\theta, \widehat{\theta}_{b,n}) \le d\log(1 + 2M/m) + 2\log\frac{1}{\delta}\right\} .$$

OTCS for Oblivious Design

Let $\mathcal{S}_{n,b} = \{\theta : \max_{t\in[n]} |\langle \theta, X_t\rangle| \le b\}$, let $\widehat{\theta}_{n,b} = \operatorname{argmin}_{\theta\in\mathcal{S}_{n,b}} \sum_{t=1}^n \ell_t(\theta)$ and let $\Psi(\theta) = \sum_{t=1}^n \psi(\langle \theta, X_t\rangle)$. Suppose that $\theta^\star$ satisfies $\sup_{t\in[n]} |\langle \theta^\star, X_t\rangle| \le b$ (w.p. 1) that ψ is M-smooth on $\mathbb{R}$ and m-strongly-convex on $[-b, b]$. Set $\rho(\theta) = \frac{1}{2\gamma^2}\|\theta - \theta^\star\|^2_{\Lambda_n}$.

For any $\delta \in (0, 1]$, the OTCS with oblivious design tells us that

$$\mathbb{P}\left(\sum_{t=1}^n \ell_t(\theta^\star) - \sum_{t=1}^n \ell_t(\widehat{\theta}_{n,b}) \le \frac{d}{2}\log(1 + M\gamma^2) + \frac{1}{2\gamma^2}\|\widehat{\theta}_{n,b} - \theta^\star\|^2_{\Lambda_n} + \log\frac{1}{\delta}\right) .$$

Using the first-order optimality condition satisfied by $\widehat{\theta}_{n,b}$ and strong convexity of ψ on $[-b, b]$,

$$\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) \le \sum_{t=1}^n \ell_t(\theta) - \sum_{t=1}^n \ell_t(\widehat{\theta}_{n,b}) , \quad \text{and} \quad \frac{1}{2\gamma^2}\|\widehat{\theta}_{n,b} - \theta^\star\|^2_{\Lambda_n} \le \frac{1}{m\gamma^2}\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) .$$

Therefore (with $\gamma^2 = 2/m$), the set Θ_n satisfies $\mathbb{P}(\theta^\star \notin \Theta_n) \le \delta$, where

$$\Theta_n = \left\{\theta : \mathcal{B}_\Psi(\theta, \widehat{\theta}_{b,n}) \le d\log(1 + 2M/m) + 2\log\frac{1}{\delta}\right\} .$$

For linear models, we get the ellipsoid

$$\Theta_n = \left\{\theta : \|\theta - \widehat{\theta}_{b,n}\|^2_{\Lambda_n} \le 2d\log(3) + 4\log\frac{1}{\delta}\right\} .$$

Conclusion

Conclusion

- Regret bounds can be converted into confidence sets/sequences for GLMs

Conclusion

- Regret bounds can be converted into confidence sets/sequences for GLMs
- Also in the paper: confidence sets with different shapes, confidence sets for sparse GLMs

Conclusion

- Regret bounds can be converted into confidence sets/sequences for GLMs
- Also in the paper: confidence sets with different shapes, confidence sets for sparse GLMs
- How to do this with the Normalised Maximum Likelihood forecaster instead of EWA?

Conclusion

- Regret bounds can be converted into confidence sets/sequences for GLMs
- Also in the paper: confidence sets with different shapes, confidence sets for sparse GLMs
- How to do this with the Normalised Maximum Likelihood forecaster instead of EWA?
- For which q^n (if any) is $\exp\left(\sum_{t=1}^{n}\left(\ell_t(\theta^\star) - \mathcal{L}_t(q_t)\right)\right)$ an optimal e-variable/process?

Conclusion

- Regret bounds can be converted into confidence sets/sequences for GLMs
- Also in the paper: confidence sets with different shapes, confidence sets for sparse GLMs
- How to do this with the Normalised Maximum Likelihood forecaster instead of EWA?
- For which q^n (if any) is $\exp\left(\sum_{t=1}^{n}\left(\ell_t(\theta^\star) - \mathcal{L}_t(q_t)\right)\right)$ an optimal e-variable/process?

The end. Thank you!