From Convex Analysis to Learning, Prediction, and Elicitation*
Lecture 7: Convex Conjugation and Fenchel-Young Divergence

Lunjia Hu

We have seen the Follow-the-Regularized-Leader (FTRL) algorithm for Online Convex Optimization (OCO). Given learner's domain X, adversary's strategy space $F \subseteq \{\text{all functions } f : X \to \mathbb{R}\}$, regularizer $\varphi : X \to \mathbb{R}$ and learning rate $\eta > 0$, FTRL is the following algorithm:

- Initialize $g_1(x) = 0$ for every $x \in X$;
- In each round $t = 1, \ldots, T$,
  1. play
  $$x_t \leftarrow \arg\min_{x \in X}(\varphi(x) + g_t(x)), \tag{1}$$
  2. observe $f_t \in F$, and
  3. update $g_{t+1} \leftarrow g_t + \eta f_t$.

We have shown that FTRL achieves the following regret bound: for every $x^* \in X$,

$$\begin{aligned}\eta \cdot \sum_{t=1}^{T}(f_t(x_t) - f_t(x^*)) &= \Gamma_\varphi(x^*, g_1) - \Gamma_\varphi(x^*, g_{T+1}) + \sum_{t=1}^{T}\Gamma_\varphi(x_t, g_{t+1}) \\ &\le \Gamma_\varphi(x^*, g_1) + \sum_{t=1}^{T}\Gamma_\varphi(x_t, g_{t+1}). \end{aligned} \tag{2}$$

This regret bound is stated using the notion of divergence Γ_φ we defined in the previous lecture.

A special case of OCO is the experts problem, where $X = \Delta_d$ and F consists of linear functions $f(x) = \langle x, y\rangle$ for $y \in [-1, 1]^d$. We claimed in the previous lecture that for this problem, a good choice of the regularizer φ is the negative Shannon entropy:

$$\varphi(x) := \sum_{i=1}^{d} x_i \ln x_i \in [-\ln d, 0] \quad \text{for every } x = (x_1, \ldots, x_d) \in \Delta_d. \tag{3}$$

This choice bounds the right-hand side of (2) as follows:

$$\Gamma_\varphi(x^*, g_1) \le \ln d, \tag{4}$$

$$\Gamma_\varphi(x_t, g_{t+1}) \le \eta^2/2. \tag{5}$$

*https://lunjiahu.com/convex-analysis/

Plugging these into (2) and picking the optimal choice of $\eta = \sqrt{\frac{2 \ln d}{T}}$, we get

$$\sum_{t=1}^{T} (f_t(x_t) - f_t(x^*)) \leq \frac{\ln d}{\eta} + \frac{T\eta}{2} = \sqrt{2T \ln d}.$$

In this lecture, we will dive deep into the reasoning behind the regularizer choice (3) and better understand the meaning of the two divergence bounds (4) and (5). To achieve this goal, we need some basic knowledge about convex analysis.

1 Subgradient

Definition 1 (subgradient)**.** *Let $X \subseteq \mathbb{R}^d$ be a non-empty set. Let $\varphi : X \to \mathbb{R}$ be a function on X. For $x \in X$ and $y \in \mathbb{R}^d$, we say y is a* subgradient *of φ at x if*

$$\varphi(x') - \varphi(x) \geq \langle x' - x, y \rangle \quad \textit{for every } x' \in X,$$

or equivalently,

$$\langle x, y \rangle - \varphi(x) \geq \langle x', y \rangle - \varphi(x') \quad \textit{for every } x' \in X,$$

or equivalently,

$$\langle x, y \rangle - \varphi(x) = \max_{x' \in X} (\langle x', y \rangle - \varphi(x')). \tag{6}$$

Theorem 1. *Let $X \subseteq \mathbb{R}^d$ be an open convex set and let $\varphi : X \to \mathbb{R}$ be a convex function. For every $x \in X$, there exists a subgradient $y \in \mathbb{R}^d$ of φ at x.*

Proof. Define $S := \{(x, z) \in X \times \mathbb{R} : z > \varphi(x)\}$. Since φ is a convex function on a convex domain X, it is easy to verify that S is a convex set. For every $x \in X$, it is clear that $(x, \varphi(x)) \notin S$, so by the hyperplane separation theorem, there exists $h = (y, v) \in \mathbb{R}^d \times \mathbb{R}$ such that $h \neq \mathbf{0}$ and

$$\langle x, y \rangle + \varphi(x)v \geq \langle x', y \rangle + z'v \quad \text{for every } (x', z') \in S. \tag{7}$$

For every $(x', z') \in S$, we can arbitrarily increase z' and result still belongs to S. Thus (7) can hold only when $v \leq 0$. It is also easy to show that $v \neq 0$ by contradiction. Indeed, if $v = 0$, then (7) implies $\langle x, y \rangle \geq \langle x', y \rangle$ for some $y \neq \mathbf{0}$ and every $x' \in X$, contradicting the assumption that x belongs to the open set X.

We have now shown that $v < 0$ must hold. By scaling all coordinates of h with the same positive factor, we can assume without loss of generality that $v = -1$. Plugging it into (7) and taking the limit $z' \to \varphi(x')$, we get

$$\langle x, y \rangle - \varphi(x) \geq \langle x', y \rangle - \varphi(x') \quad \text{for every } x' \in X.$$

This completes the proof that y is a subgradient of φ at x. □

2 Convex Conjugation

Definition 2 (Convex Conjugation (Legendre Transformation)). *Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be an arbitrary function. We define the* convex conjugate *of φ as the function $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ where*

$$\psi(y) := \sup_{x \in \mathbb{R}^d} (\langle x, y\rangle - \varphi(x)) \quad \text{for every } y \in \mathbb{R}^d.$$

Lemma 2. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and let $\|\cdot\|_*$ be its dual norm. Then the convex conjugate of $\varphi(x) = \frac{1}{2}\|x\|^2$ is $\psi(y) = \frac{1}{2}\|y\|_*^2$.*

Lemma 3 (Inverse monotonicity of convex conjugation). *Let $\varphi_1, \varphi_2 : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be arbitrary functions, and let ψ_1, ψ_2 be their convex conjugates, respectively. Assume $\varphi_1(x) \geq \varphi_2(x)$ for every $x \in \mathbb{R}^d$. Then $\psi_1(y) \leq \psi_2(y)$ for every $y \in \mathbb{R}^d$.*

Definition 3 (Closed convex function). *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be an arbitrary function. Its* epigraph *is defined as*

$$E_f := \{(x, z) \in \mathbb{R}^d \times \mathbb{R} : z \geq f(x)\}.$$

We say f is a closed convex function *if E_f is a closed convex set.*

Lemma 4. *Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be an arbitrary function and let ψ be its convex conjugate. Then ψ is a closed convex function.*

Proof. The epigraph E_ψ of ψ can be expressed as follows:

$$E_\psi = \{(y, z) \in \mathbb{R}^d \times \mathbb{R} : z \geq \langle x, y\rangle - \varphi(x) \text{ for every } x \in \mathbb{R}^d\}.$$

This means that $E_\psi = \bigcap_{x \in \mathbb{R}^d} S_x$, where

$$S_x := \{(y, z) \in \mathbb{R}^d \times \mathbb{R} : z \geq \langle x, y\rangle - \varphi(x)\}.$$

It is easy to verify that each S_x is a closed convex set (in fact, it is a half-space, the whole space $\mathbb{R}^d$, or the empty set $\emptyset$), so $E_\psi = \bigcap_{x \in \mathbb{R}^d} S_x$ is also a closed convex set. □

Theorem 5 (Double conjugation). *Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a closed convex function and let ψ be its convex conjugate. Then φ is also the convex conjugate of ψ.*

Proof. If $\varphi(x) = +\infty$ for every $x \in \mathbb{R}^d$, then $\psi(y) = -\infty$ for every $y \in \mathbb{R}^d$. The theorem holds trivially in this case. We thus assume $\varphi(x_0) \in \mathbb{R}$ for some $x_0 \in \mathbb{R}^d$.

By the assumption that ψ is the convex conjugate of φ, we have

$$\psi(y) \geq \langle x, y\rangle - \varphi(x) \quad \text{for every } x, y \in \mathbb{R}^d.$$

Therefore,

$$\varphi(x) \geq \langle x, y\rangle - \psi(y) \quad \text{for every } x, y \in \mathbb{R}^d.$$

This means that $\varphi(x) \geq \sup_{y \in \mathbb{R}^d}(\langle x, y\rangle - \psi(y))$ for every $x \in \mathbb{R}^d$. It remains to prove the reverse inequality

$$\varphi(x) \leq \sup_{y \in \mathbb{R}^d} (\langle x, y\rangle - \psi(y)) \quad \text{for every } x \in \mathbb{R}^d. \tag{8}$$

To prove (8), it suffices to prove that for every $x \in \mathbb{R}^d$ and every $z < \varphi(x)$, it holds that

$$z \leq \sup_{y \in \mathbb{R}^d} (\langle x, y \rangle - \psi(y)).$$

Since $z < \varphi(x)$, we have $(x, z) \notin E_\varphi$. Since E_φ is a closed convex set, we have strict hyperplane separation: there exists $h = (y, v) \in \mathbb{R}^d \times \mathbb{R}$ and $\varepsilon > 0$ such that

$$\langle x, y \rangle + zv > \langle x', y \rangle + z'v + \varepsilon \quad \text{for every } (x', z') \in E_\varphi. \tag{9}$$

For every $(x', z') \in E_\varphi$, we can arbitrarily increase z' and the result still belongs to E_φ. Thus (9) holds only when $v \leq 0$. We divide the rest of the proof into two cases.

Case 1: $\varphi(x) \in \mathbb{R}$. We show that $v < 0$ must hold. Assume by contradiction that $v = 0$. Then (9) implies $\langle x, y \rangle > \langle x', y \rangle$ for every $(x', z') \in E_\varphi$. Since $\varphi(x) \in \mathbb{R}$, we have $(x, \varphi(x)) \in E_\varphi$. Choosing $(x', z') = (x, \varphi(x)) \in E_\varphi$, we get $\langle x, y \rangle > \langle x, y \rangle$, a contradiction.

We have now shown $v < 0$. By scaling all coordinates of $h = (y, v)$ with the same positive constant, we can assume without loss of generality that $v = -1$. Taking $z' \to \varphi(x')$ in (9), we get

$$\langle x, y \rangle - z \geq \sup_{x' \in \mathbb{R}^d} (\langle x', y \rangle - \varphi(x')) = \psi(y).$$

Therefore,

$$z \leq \langle x, y \rangle - \psi(y) \leq \sup_{y \in \mathbb{R}^d} (\langle x, y \rangle - \psi(y)).$$

Case 2: $\varphi(x) = +\infty$. Recall our assumption that $\varphi(x_0) \in \mathbb{R}$ for some $x_0 \in \mathbb{R}^d$. Pick an arbitrary $z_0 \in \mathbb{R}$ such that $z_0 < \varphi(x_0)$. By our analysis of Case 1, there exists $(y_0, v_0) \in \mathbb{R}^d \times \mathbb{R}$ such that $v_0 < 0$ and

$$\langle x_0, y_0 \rangle + z_0 v_0 > \langle x', y_0 \rangle + z'v_0 \quad \text{for every } (x', z') \in E_\varphi. \tag{10}$$

For $\alpha \geq 0$, define $y' := y + \alpha y_0$ and $v' := v + \alpha v_0$. Combining (9) and (10), we have

$$\langle x, y \rangle + zv + \alpha(\langle x_0, y_0 \rangle + z_0 v_0) > \langle x', y' \rangle + z'v' + \varepsilon \quad \text{for every } (x', z') \in E_\varphi. \tag{11}$$

When $\alpha = 0$, we have

$$\langle x, y \rangle + zv + \alpha(\langle x_0, y_0 \rangle + z_0 v_0) = \langle x, y' \rangle + zv'.$$

By the continuity of both sides as functions of α, for every sufficiently small $\alpha > 0$, we have

$$\langle x, y \rangle + zv + \alpha(\langle x_0, y_0 \rangle + z_0 v_0) < \langle x, y' \rangle + zv' + \varepsilon. \tag{12}$$

Combining (11) and (12), we get

$$\langle x, y' \rangle + zv' > \langle x', y' \rangle + z'v' \quad \text{for every } (x', z') \in E_\varphi. \tag{13}$$

Since $v' = v + \alpha v_0$ where $v \leq 0$ and $v_0 < 0$, we have $v' < 0$. By scaling (y', v') using a positive factor, we can assume without loss of generality that $v' = -1$ in (13). Taking $z' \to \varphi(x')$, we get

$$\langle x, y' \rangle - z \geq \sup_{x' \in \mathbb{R}^d} (\langle x', y' \rangle - \varphi(x')) = \psi(y').$$

Therefore,

$$z \leq \langle x, y' \rangle - \psi(y') \leq \sup_{y \in \mathbb{R}^d} (\langle x, y' \rangle - \psi(y')). \qquad \square$$

3 Fenchel-Young Divergence

Definition 4 (Fenchel-Young Divergence)**.** *Let $\varphi, \psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be arbitrary functions. For $x, y \in \mathbb{R}^d$, their* Fenchel-Young divergence *is defined as*

$$\Gamma_{\varphi,\psi}(x, y) := \varphi(x) + \psi(y) - \langle x, y\rangle \in \mathbb{R} \cup \{+\infty\}.$$

We say a function $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is *proper* if there exists some $x_0 \in \mathbb{R}^d$ such that $f(x_0) \in \mathbb{R}$ (i.e. $f(x_0) < +\infty$). It is easy to show that if φ is a proper function, its convex conjugate ψ satisfies $\psi(y) > -\infty$ for every $y \in \mathbb{R}^d$.

Theorem 6. *Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be an arbitrary proper function. Let $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be the convex conjugate of φ. Then for every $x, y \in \mathbb{R}^d$,*

$$\begin{aligned}\Gamma_{\varphi,\psi}(x, y) &= \sup_{x' \in X} (\langle x', y\rangle - \varphi(x')) - (\langle x, y\rangle - \varphi(x)) \\ &= (\varphi(x) - \langle x, y\rangle) - \inf_{x' \in X} (\varphi(x') - \langle x', y\rangle) \qquad (14) \\ &\geq 0.\end{aligned}$$

Moreover, the following two statements are equivalent:

1. *$\Gamma_{\varphi,\psi}(x, y) = 0$;*
2. *y is a subgradient of φ at x.*

Lemma 7. *Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be an arbitrary proper function. Let $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be the convex conjugate of φ. For a fixed pair $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ satisfying $\Gamma_{\varphi,\psi}(x, y) = 0$, define functions $f, g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ as follows:*

$$\begin{aligned} f(x') &:= \Gamma_{\varphi,\psi}(x + x', y) \quad \text{for every } x' \in \mathbb{R}^d, \\ g(y') &:= \Gamma_{\varphi,\psi}(x, y + y') \quad \text{for every } y' \in \mathbb{R}^d. \end{aligned}$$

Then g is the convex conjugate of f.

Proof. For every $y' \in \mathbb{R}^d$,

$$\begin{aligned} &\sup_{x' \in \mathbb{R}^d} (\langle x', y'\rangle - f(x')) \\ =\ &\sup_{x' \in \mathbb{R}^d} (\langle x', y'\rangle - \varphi(x + x') - \psi(y) + \langle x + x', y\rangle) \\ =\ &\sup_{x' \in \mathbb{R}^d} (\langle x' + x, y' + y\rangle - \varphi(x + x')) - \psi(y) - \langle x, y'\rangle \\ =\ &\psi(y + y') - \psi(y) - \langle x, y'\rangle \\ =\ &\psi(y + y') + \varphi(x) - \langle x, y + y'\rangle \qquad (\text{because } \varphi(x) + \psi(y) - \langle x, y\rangle = \Gamma_{\varphi,\psi}(x, y) = 0) \\ =\ &\Gamma_{\varphi,\psi}(x, y + y') \\ =\ &g(y'). \end{aligned}$$

□

4 Strong Convexity and Smoothness

Definition 5 (Strong convexity)**.** *Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function and let $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be its convex conjugate. For $\lambda \geq 0$, we say φ is λ-strongly convex w.r.t. norm $\|\cdot\|$ if for every pair $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ satisfying $\Gamma_{\varphi,\psi}(x, y) = 0$, it holds that*

$$\Gamma_{\varphi,\psi}(x', y) \geq \frac{\lambda}{2}\|x' - x\|^2 \quad \textit{for every } x' \in \mathbb{R}^d,$$

or equivalently,

$$\varphi(x') - \varphi(x) - \langle x' - x, y\rangle \geq \frac{\lambda}{2}\|x' - x\|^2 \quad \textit{for every } x' \in \mathbb{R}^d.$$

The definition of smoothness changes the "$\geq$" signs in Definition 5 to "$\leq$" signs:

Definition 6 (Smoothness)**.** *Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function and let $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be its convex conjugate. For $\lambda \geq 0$, we say φ is λ-smooth w.r.t. norm $\|\cdot\|$ if for every pair $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ satisfying $\Gamma_{\varphi,\psi}(x, y) = 0$, it holds that*

$$\Gamma_{\varphi,\psi}(x', y) \leq \frac{\lambda}{2}\|x' - x\|^2 \quad \textit{for every } x' \in \mathbb{R}^d,$$

or equivalently,

$$\varphi(x') - \varphi(x) - \langle x' - x, y\rangle \leq \frac{\lambda}{2}\|x' - x\|^2 \quad \textit{for every } x' \in \mathbb{R}^d. \tag{15}$$

If a function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is λ-smooth, then by (15), it must hold that $\varphi(x) < +\infty$ for every $x \in \mathbb{R}^d$.

Theorem 8. *Let $\varphi, \psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a pair of mutually conjugate functions. Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and let $\|\cdot\|_*$ be its dual norm. Then for every $\lambda > 0$, the following two statements are equivalent:*

1. *φ is λ-strongly convex w.r.t. $\|\cdot\|$;*
2. *ψ is $(1/\lambda)$-smooth w.r.t. $\|\cdot\|_*$.*

Proof. By Definition 5, the statement that φ is λ-strongly convex w.r.t. $\|\cdot\|$ is equivalent to the following statement: for every pair $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ satisfying $\Gamma_{\varphi,\psi}(x, y) = 0$, it holds that

$$\Gamma_{\varphi,\psi}(x + x', y) \geq \frac{\lambda}{2}\|x'\|^2 \quad \text{for every } x' \in \mathbb{R}^d. \tag{16}$$

Similarly, by Definition 6, the statement that ψ is $(1/\lambda)$-smooth w.r.t. $\|\cdot\|_*$ is equivalent to the following statement: for every pair $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ satisfying $\Gamma_{\varphi,\psi}(x, y) = 0$, it holds that

$$\Gamma_{\varphi,\psi}(x, y + y') \leq \frac{1}{2\lambda}\|y'\|_*^2 \quad \text{for every } y' \in \mathbb{R}^d. \tag{17}$$

By Lemma 2, the functions $\frac{\lambda}{2}\|x'\|^2$ and $\frac{1}{2\lambda}\|y'\|_*^2$ are conjugate functions of each other. By Lemma 7, the functions $\Gamma_{\varphi,\psi}(x + x', y)$ and $\Gamma_{\varphi,\psi}(x, y + y')$ are also conjugate functions of each other. Thus Lemma 3 implies that (16) and (17) are equivalent. □

5 Back to the Experts Problem

Now we explain what exact properties of φ lead to the two divergence bound (4) and (5). We first show that (5) comes from the strong convexity of φ:

Theorem 9. *The negative Shannon entropy φ in (3) is 1-strongly convex in the ℓ_1 norm $\|\cdot\|_1$.*

Theorem 9 is the famous *Pinsker's Inequality*. We omit the proof here.

Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be the convex conjugate of the negative Shannon entropy φ in (3). By Theorems 8 and 9, we know that ψ is 1-smooth w.r.t. the ℓ_∞ norm $\|\cdot\|_\infty$. This proves the following lemma, where an equivalent form of the lemma appeared in the previous lecture:

Lemma 10. *Let $\varphi : \Delta_d \to \mathbb{R}$ be the negative Shannon entropy (3), and let $\psi : \mathbb{R}^d \to \mathbb{R}$ be its convex conjugate. Let $(x, z) \in \Delta_d \times \mathbb{R}^d$ be a pair satisfying $\Gamma_{\varphi,\psi}(x, z) = 0$. Then*

$$\Gamma_{\varphi,\psi}(x, z') \le \frac{1}{2}\|z - z'\|_\infty^2 \quad \textit{for every } z' \in \mathbb{R}^d.$$

When we apply FTRL to the experts problem, the functions g_t are linear functions of $x \in \Delta_d$. We can write $g_t(x) = -\langle x, z_t\rangle$ for some $z_t \in \mathbb{R}^d$. By (14), we have

$$\begin{aligned}\Gamma_\varphi(x, g_t) &= \varphi(x) + g_t(x) - \inf_{x' \in \Delta_d}(\varphi(x') + g_t(x')) \\ &= \varphi(x) - \langle x, z_t\rangle - \inf_{x' \in \Delta_d}(\varphi(x') - \langle x', z_t\rangle) \\ &= \Gamma_{\varphi,\psi}(x, z_t).\end{aligned}$$

In FTRL, we have $\Gamma_{\varphi,\psi}(x_t, z_t) = \Gamma_\varphi(x_t, g_t) = 0$. Thus the divergence bound (5) follows from Lemma 10:

$$\Gamma_\varphi(x_t, g_{t+1}) = \Gamma_{\varphi,\psi}(x_t, z_{t+1}) \le \frac{1}{2}\|z_{t+1} - z_t\|_\infty^2 \le \eta^2/2.$$

The last inequality holds because in FTRL, we have $z_{t+1} = z_t - \eta y_t$ for some $y_t \in [-1, 1]^d$.

We have shown that (5) comes from the strong convexity of the regularizer φ. Now we show that (4) comes from the boundedness of φ. Since g_1 is the constant zero function,

$$\Gamma_\varphi(x^*, g_1) = \varphi(x^*) - \inf_{x' \in \Delta_d}\varphi(x') \le \sup_{x' \in \Delta_d}\varphi(x') - \inf_{x' \in \Delta_d}\varphi(x') = \ln d.$$

In summary, the two divergence bounds (4) and (5) hold because φ is both bounded and strongly convex. In general, if we choose an arbitrary regularizer $\varphi : \Delta_d \to \mathbb{R}$ that is λ-strongly convex and M-bounded: $\sup_{x' \in \Delta_d}\varphi(x') - \inf_{x' \in \Delta_d}\varphi(x') \le M$, then we have

$$\begin{aligned}\Gamma_\varphi(x^*, g_1) &\le M, \\ \Gamma_\varphi(x_t, g_{t+1}) &\le \eta^2/(2\lambda).\end{aligned}$$

The corresponding regret bound we get for the optimal choice of $\eta = \sqrt{\frac{2\lambda M}{T}}$ is

$$\text{regret} \le \frac{M}{\eta} + \frac{\eta T}{2\lambda} = \sqrt{\frac{2MT}{\lambda}}.$$

To get the best regret bound, we would like to make M/λ as small as possible. The negative Shannon entropy is a good regularizer exactly because it makes M/λ small.